Sixty Alleles of the *ALS7* Open Reading Frame in *Candida albicans*: *ALS7* Is a Hypermutable Contingency Locus

Ningxin Zhang, Annette L. Harrex, Barbara R. Holland, et al.

Genome Res. 2003 13: 2005-2017
Access the most recent version at doi:10.1101/gr.1024903

Article

Sixty Alleles of the *ALS7* Open Reading Frame in *Candida albicans*: *ALS7* Is a Hypermutable Contingency Locus

Ningxin Zhang,[1] Annette L. Harrex,[3] Barbara R. Holland,[2,4] Lauren E. Fenton,[3] Richard D. Cannon,[3] and Jan Schmid[1,5]

[1]*Institute of Molecular BioSciences,* [2]*Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand;* [3]*Department of Oral Sciences and Orthodontics, University of Otago, Dunedin, New Zealand*

The *ALS* (agglutinin-like sequence) gene family encodes proteins that play a role in adherence of the yeast *Candida albicans* to endothelial and epithelial cells. The proteins are proposed as virulence factors for this important fungal pathogen of humans. We analyzed 66 *C. albicans* strains, representing a worldwide collection of 266 infection-causing isolates, and discovered 60 alleles of the *ALS7* open reading frame (ORF). Differences between alleles were largely caused by rearrangements of repeat elements in the so-called tandem repeat domain (21 different types occurred) and the VASES region (19 different types). *C. albicans* is diploid, and combinations of *ALS7* alleles generated 49 different genotypes. *ALS7* expression was detected in samples isolated directly from five oral candidosis patients. ORFs in the opposite direction contained within the *ALS7* ORF were also transcribed in all strains tested. Isolates representing a more pathogenic general-purpose genotype (GPG) cluster of strains tended to have more tandem repeats than other strains. Two types of VASES regions were largely exclusive to GPG strains; the remaining types were largely exclusive to noncluster strains. Our results provide evidence that *ALS7* is a hypermutable contingency locus and important for the success of *C. albicans* as an opportunistic pathogen of humans.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession numbers AY170875–AY170888.]

Candida albicans is a diploid yeast with a predominantly clonal mode of reproduction (Tibayrenc 1997) that is an important opportunistic pathogen of humans (Odds 1988). We recently presented evidence of a ubiquitous general-purpose genotype (GPG) cluster of *C. albicans* strains that are exceptionally successful as human pathogens (Schmid et al. 1999). Two other studies have since provided evidence that this genotype (as deduced from the position of GPG reference strains included in the studies) was the most frequent among the infection-causing isolates they analyzed (Botterel et al. 2001; Bougnoux et al. 2002). We are currently undertaking comparisons between GPG cluster strains and other strains to identify physiological and genetic traits which are associated with their increased pathogenicity, as a means to identify pathogenicity factors relevant to interactions of *C. albicans* with the human host (Schmid et al. 1995; Giblin et al. 2001).

During a screen for genomic differences between GPG cluster strains and other strains using amplified fragment length polymorphism (AFLP) analysis, we discovered that the open reading frame (ORF) of the *ALS7* gene (Hoyer and Hecht 2000) showed GPG cluster-specific sequence modifications. The gene is part of a family of large agglutinin-like cell surface glycoproteins that mediate adherence of *C. albicans* to host tissue (Gaur and Klotz 1997; Chandra et al. 2001; Hoyer 2001; Fu et al. 2002). *ALS* genes are also present in *Candida dubliniensis* and *Candida tropicalis* (Hoyer et al. 2001). *C. albicans ALS* genes have been shown to be differentially expressed in a strain-dependent manner (Hoyer et al. 1998a) in response to parameters such as growth stage (Hoyer et al. 1998b), morphological transition between yeast and hyphal growth (Hoyer et al. 1998a; Murad et al. 2001), and biofilm formation (Chandra et al. 2001). It has also been documented that alterations in repeat regions of *ALS* genes lead to strain-specific variability in the Als proteins (Hoyer et al. 1995). Taken together these observations indicate that Als proteins provide *C. albicans* with a large and flexible repertoire of similar but nonidentical surface proteins. In other eukaryotic microbial pathogens such as *Plasmodium*, *Trypanosoma*, and *Giardia*, such a repertoire of proteins is considered important not only for adhesion but also for evasion of host defenses (Nash 1997; Ramasamy 1998; Barry and McCulloch 2001). High mutation rates in the genes encoding these proteins are a crucial part of generating diversity. The term "hypermutable contingency genes" was coined for such genes (Duncan et al. 1991; Moxon and Thaler 1997; Barry and McCulloch 2001; Bridges 2001). Aside from encoding cell surface proteins, hypermutable contingency genes are defined by (1) being potentially highly mutable, and (2) mutations which are not caused by error-prone DNA replication but are rather specific, being brought about by recombination (or methylation) events, involving oligonucleotide repeats or homoploymeric tracts (Moxon and Thaler 1997; Bridges 2001). *ALS7* would be particularly well suited as a hypermutable contingency gene in *C. albicans*; its ORF contains not only the tandem repeat domain, characteristic of all *ALS* genes, but in addition a second repeat region, the so-called VASES region.

We present here evidence, based on the allelic variation at the *ALS7* locus in a large set of strains representing the major phylogenetic lineages in a worldwide collection of 266 infection-

[4]Present address: Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.
[5]Corresponding author.
E-MAIL J.Schmid@massey.ac.nz; FAX 64 (6) 350-5688.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1024903.

causing isolates, that *ALS7* is a hypermutable contingency locus. We also describe restraints on the rearrangements of repeat units in the *ALS7* ORF and expression data which allow conclusions to be drawn regarding the role of *ALS7* in the pathogenesis of candidosis in humans.

RESULTS

An AFLP Marks Alterations in the *ALS7* Gene Specific to GPG Cluster Strains

Following the discovery that a ubiquitous GPG cluster of *C. albicans* causes candidosis 10–100 times more often than other clusters of strains (Schmid et al. 1999), we used AFLP analysis (Vos et al. 1995) for genomic comparisons between GPG strains and other strains. We searched for AFLPs highly specific to GPG strains as markers of loci contributing to pathogenicity. In these comparisons we used two partially overlapping sets of isolates, each set containing 25 GPG cluster strains and 16 strains not belonging to the cluster (noncluster strains), representing a worldwide collection of 266 infection-causing isolates (Schmid et al. 1999). A list of strains in the sets and their *ALS7* alleles can be found in the Supplemental Material (available online at www.genome.org). Set 1 contained isolates chosen on the basis that they best represent features that distinguish GPG cluster and noncluster strains (see Methods for details). Set 2 contained isolates that best represent the entire spectrum of GPG cluster strains and the entire spectrum of noncluster strains, respectively (this includes borderline strains with an uncertain affiliation; see Methods for details). Laboratory strain SC5314, used to construct the Stanford genome database (http://www-sequence.stanford.edu:8080/bin/blastncontigs6), was also included in both sets, as a GPG cluster strain (Giblin et al. 2001), bringing the number of GPG model strains to 26.

In a comparison between GPG cluster and noncluster isolates in Set 1, we discovered a 947-bp AFLP fragment, MC5-c, present in 21 of 26 GPG cluster strains (81%) and in one of 16 noncluster strains (6%). The fragment was also present in 22 of 26 GPG cluster strains in Set 2 (85%) and three of 16 (19%) noncluster strains in Set 2. AFLP products from two GPG cluster strains were cloned and sequenced. Database searches showed homology to part of the *ALS7* gene in strain 1161 (accession no. AF201684; Hoyer and Hecht 2000) and SC5314 (bp 7513–1511 in contig 6–2514; http://www-sequence.stanford.edu:8080/contigs/Contig6-2514).

Sixty Alleles of the *ALS7* Gene and 49 *ALS7* Genotypes in 66 *C. albicans* Strains

According to the definition of Hoyer (2001), the *ALS7* gene contains three domains (Fig. 1), and two of these, named the tandem repeat domain and the 3′domain, contain repeats; the repeats in the 3′ domain are contained in the so-called VASES region. The GPG cluster-specific AFLP fragment overlapped with the VASES region, and amplification of the region of *ALS7* corresponding to the AFLP fragment from noncluster strains gave a variety of PCR products. This suggested that the repeat-containing parts of the *ALS7* ORF might undergo frequent rearrangements typical for hypermutable contingency loci, as outlined in the Introduction. We therefore decided to assess variability in both repeat regions for all of our model strains.

The tandem repeat domain is composed of 108-bp units (Hoyer and Hecht 2000; Hoyer 2001). We amplified the entire tandem repeat domain in different strains and used the size of the amplicons to determine the number of repeat units present (Fig. 2A). We used restriction digestion of the amplicons to verify that the PCR products were indeed composed of 108-bp units (Fig. 2B). The analysis of the VASES region was more complicated. Although the region is named after a 15-bp repeat encoding the amino acid sequence VASES (in addition there is a similar version encoding VTSES), it contains three other repeat elements. We found that the region is usually assembled from one or more units, each comprising an 87-bp repeat, followed by a 138-bp repeat, followed by a stretch of VA/TSES repeats, followed by a 141-bp repeat. At the end of the region is one additional 87-bp repeat (Fig. 1). We characterized the VASES regions by first amplifying the entire region (Fig. 2C) and then, using this PCR product as a template, carrying out additional amplifications to deduce arrangement of individual units (data not shown; see Methods for details). From the product sizes we calculated the number of repeats in each unit.

Where PCR analysis indicated the presence of two different-sized tandem repeat domains or VASES regions in the same strain, we also generated PCR products spanning both the tandem domain and VASES region to derive the size of the two alleles present (data not shown). We tested the accuracy of our PCR-based derivations of arrangements of repeat units by comparing them with the published sequences available for SC5314 and 1161 and with our sequences obtained for three tandem repeat domains and 12 VASES regions (see below). In all cases, the PCR-based estimates proved correct.

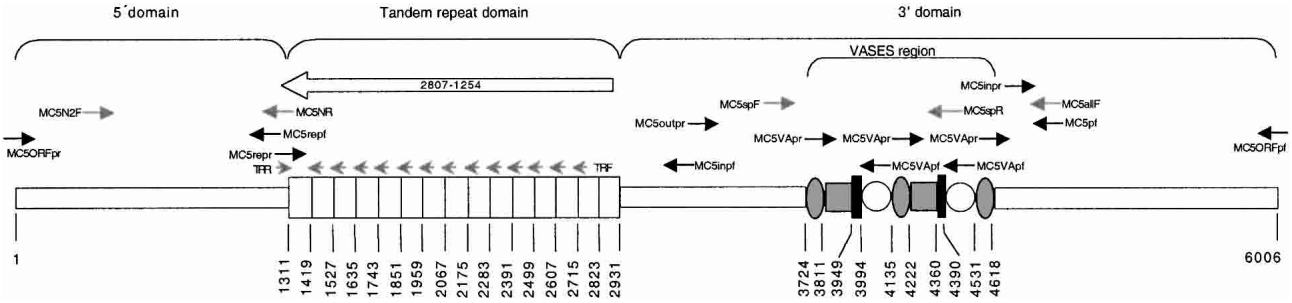


Figure 1 Schematic representation of the *ALS7* gene. The schematic is based on the gene in strain SC5314. Shapes are used to show the arrangement of the following repetitive elements: the 108-bp element (hollow boxes); the 87-bp element (gray ovals); the 138-bp element (gray boxes); VA/TSES repeats (black boxes; the thickness of the box varies according to the number of tandemly arranged 15-bp units present); and the 141-bp element (hollow circles). At the top of the figure, brackets show the borders of the different domains and regions within the gene, as defined by Hoyer (2001). Labeled thin arrows represent the primers used, with gray arrows indicating primers used for RT–PCR. The hollow arrow indicates a reverse ORF. Numbers at the bottom indicate the position within the gene (in base pairs). Note that the GPG-specific AFLP fragment corresponds to the region from bp 3719 to bp 4957 in the SC5314 gene; however, a point mutation in bp 3723 in SC5314 prevents amplification in the AFLP reaction. Note also that the VASES region of SC5314 is 290 bp longer than that amplified by AFLP for most cluster strains; thus SC5314 DNA did not generate the GPG cluster-specific AFLP product.

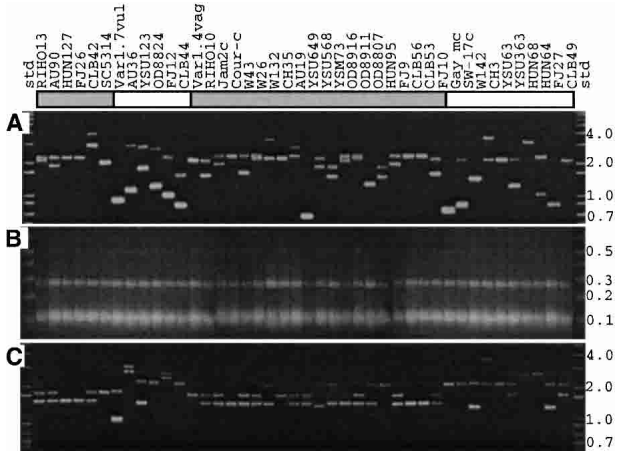


Figure 2 Characterization of *ALS7* alleles by PCR and restriction analysis in the 42 strains of model Set 1. GPG cluster strains are marked by a shaded bar under the strain name, other strains by a hollow bar. (*A*) PCR amplification of the tandem repeat domain, using primers MC5inpf and MC5repr. (*B*) *Bsm*AI digest of the products; a single *Bsm*AI site in each 108-bp repeat is predicted from the nucleotide sequence, and so digestion of the PCR product should produce a 108-bp band plus three other bands of very similar molecular weight: an 80-bp band from the 5′ flanking sequence and bands of 95 bp and 116 bp from the 3′ flanking sequence. In addition, there should be a 270-bp band from the 3′ flanking sequence. (*C*) Amplification of the VASES region, using primers MC5spF andMC5pf. Numbers at the right of the figure give fragment sizes in kb.

Our analysis revealed the presence of 60 alleles, distinguishable by PCR analysis, in our 66 strains (Fig. 3) and 49 *ALS7* genotypes (i.e., allele combinations; for a list of genotypes, see Suppl. Material). None of these alleles was identical to that described by Hoyer and coworkers (Hoyer and Hecht 2000; Hoyer 2001), bringing the total number of alleles to 61 in 67 strains. We plotted the running sum of alleles found in our analysis versus 1/(running sum) of the number of strains analyzed by us (Fig. 4) to obtain an estimate of the maximum number of alleles present in cluster and noncluster strains. The estimate is between 30 and 40 for cluster strains and ≥90 for noncluster strains. Given that there are only very few alleles which are common to both groups of strains (see below), the total number of alleles within the species is likely to exceed 100. It is important to note, however, that the PCR analysis underestimates the diversity at the locus, because the nucleotide sequences of the repeat units appearing identical in size by PCR were not always completely identical (see below).

GPG Cluster-Specific Characteristics of *ALS7* Alleles

Few *ALS7* alleles were present in both GPG cluster and noncluster strains, and two alleles, 16a201 and 17a103, accounted for approximately half of the loci in GPG cluster strains, but were rare in noncluster strains. However, a large number of alleles occurred in only one or two strains (Fig. 5A). Likewise, no GPG cluster and noncluster strain had the same *ALS7* genotype (see list of genotypes in Suppl. Material), and only two genotypes occurred at frequencies exceeding 10% in either group (17a103/16a201 in 11% of GPG cluster strains and 17a103/17a103 in 17% of GPG cluster strains).

Differences between GPG cluster strains and other strains were more apparent when the tandem repeat domain and the VASES region were analyzed separately (Fig. 5B,C). The majority (76%) of GPG cluster strains had between 14 and 17 tandem repeats. Only 25% of noncluster strains had between 14 and 17 repeats. Conversely, 50% of noncluster strains had ≤10 repeats, but only 17% of GPG cluster strains had ≤10 repeats. In the VASES region, GPG cluster strains predominantly (83% of strains) had short versions, limited to a maximum of two stretches of VA/TSES repeats. In contrast, the majority of noncluster strains (71%) had alleles with ≥3 stretches of VA/TSES units. All of these differences between GPG strains and noncluster strains were significant (z test, $P < 0.01$).

Diversity Is Generated Using Restrained Arrangements of Conserved Units

We wanted to know whether the individual repeat units in the repetitive parts of the gene contained high numbers of random mutations, or whether allelic diversity was mainly brought about by rearranging units whose sequences were largely conserved. The degree of conservation of the units' sequences would be an indication of the functional importance of the sequence information contained in the units. We therefore analyzed nucleotide substitution rates for the individual units that make up the variable regions, namely the 108-bp repeats of the tandem repeat domain and the 15-bp, 87-bp, 138-bp, and 141-bp repeats making up the VASES region. The rates were calculated by comparing all available sequences of a given unit. This included comparisons between units within the same allele and comparisons between units in different alleles. We drew on both previously published (GenBank) and genome database sequences (*C. albicans* 1161 and SC5314, respectively), plus sequence data that we generated (Table 1; the table also contains data for nonrepetitive parts of the gene). We also generated neighbor-joining trees based on the sequences obtained for each type of repeat unit. These could highlight instances where sequences of units were fully, or highly, conserved between strains or instances where sequences of units at particular positions were more conserved than at other positions (Fig. 6).

Generally speaking, all repetitive units showed higher nucleotide substitution rates than nonrepetitive parts of the gene (Table 1). The higher substitution rates were, however, caused mainly by pronounced differences between a few types of a particular repeat unit. These types could in themselves be highly conserved, that is, a given type of unit could be found in different strains and alleles, and certain types were often associated with particular positions in the repeat regions.

In the tandem repeat domain (Table 1; Fig. 6A), both the first and last 108-bp units were distinct and highly conserved (the first unit had a nucleotide substitution rate comparable to that of nonrepetitive regions in the gene). The intermediate repeats varied more, with repeat units belonging to the same allele and from similar strains (e.g., SC3514 and RIHO13, both belonging to the GPG cluster) generally grouping together. However, there were instances where distant strains had very similar or even identical units (e.g., the SC5314 units 4, 6, and 11were identical to unit 6 in strain1161).

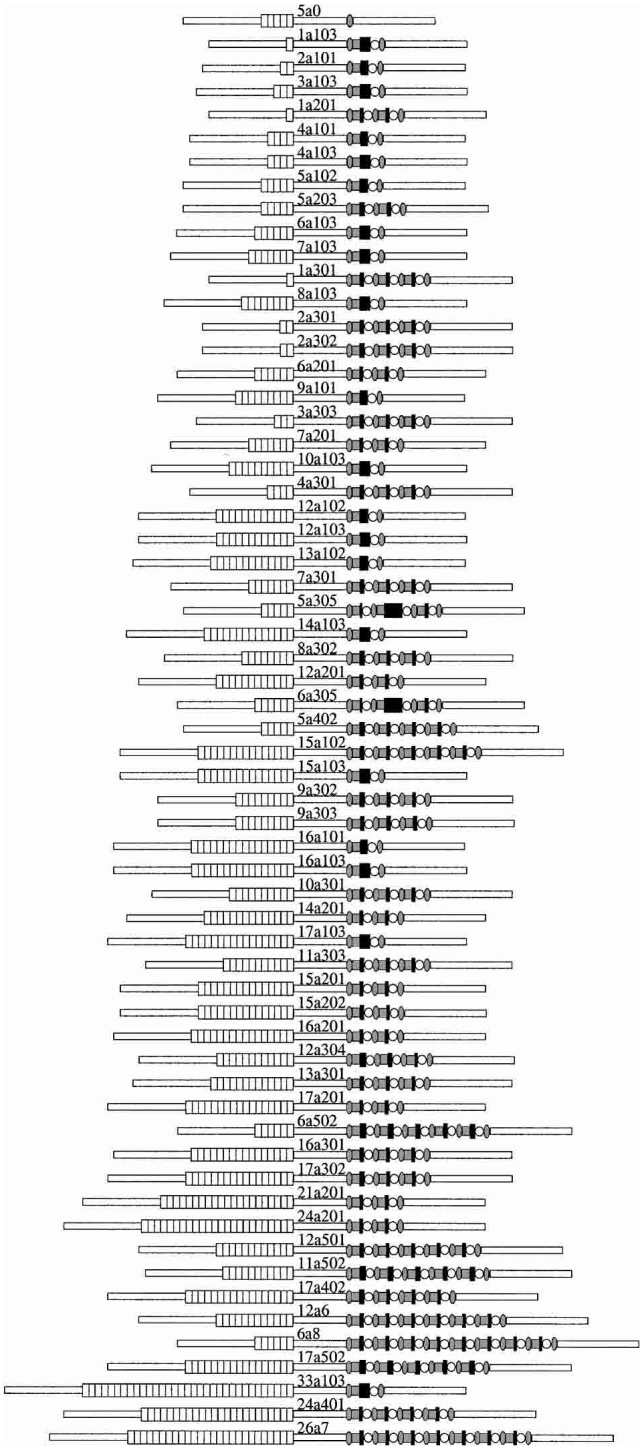
In the VASES region, there were only two types of the 15-bp repeats, encoding VASES and VTSES, after which the region is named (Table 1; Fig. 6B). Both were as conserved at the nucleic acid level as the nonrepetitive parts of the gene. In addition, each of the stretches of VA/TSES repeats comprising the VASES region started with VTSESVASES; this arrangement was thus also as conserved as the nonrepetitive parts of the gene. The probability of finding this arrangement at the beginning of all stretches if VASES and VTSES units were mixed randomly is less than 0.001 (X^2 test; the calculation took into account the frequencies with which each of the two units were encountered).

Both the first and last 87-bp units of the VASES region (Table 1; Fig. 6C) were distinct from other 87-bp units (the only excep-

tions to this rule were first and last units in laboratory strain 1161). The units had an apparent nucleotide substitution rate comparable to that of the nonrepetitive regions of the gene. Even for the intermediate 87-bp units, identical versions were present in different strains (e.g., the units from W132, OD8824, CLB44, RIHO13, and SC5314 at the bottom left of the tree in Fig. 6C).

Some versions of the 138-bp unit from distantly related strains were identical (Table 1; Fig. 6D). There are three groups containing eight, seven, and five completely identical repeats in the tree shown in Figure 6D. The nucleotide substitution rates in these groups (1/138 × number of units in group) are 0.001, 0.001, and 0.002, respectively, which is less than, or equal to, the rates determined for nonrepetitive parts of the gene.

The neighbor-joining tree for the 141-bp repeats (Fig. 6E) shows two clusters of genetically similar units predominating at the first and last positions of the VASES region; the last repeats are almost fully conserved if strain 1161 is omitted from the analysis.



The Ratio of Nonsynonymous to Synonymous Mutations in Repeat Units Indicates That They Are Under Selection

As a test of the significance of repeat regions in terms of functionality of the encoded protein, we compared repeat and nonrepetitive regions in terms of the ratio of nonsynonymous (amino acid-altering) to synonymous, (silent) mutations (d_N/d_S ratio). This ratio is held to be an indicator of natural selection (Yang and Nielsen 2002). A ratio of <1.0 indicates purifying selection; that is, amino acid substitutions are generally selected against because the vast majority will incur selective disadvantages. A ratio >1.0 indicates positive selection; that is, amino acid substitutions have a high potential of offering fitness advantages. A ratio equal to 1.0 indicates neutral evolution; that is, the absence of selection.

If repetitive parts of the ORF contribute to the functionality of the protein, their d_N/d_S ratio should be significantly different from 1.0. For the 108-bp, 87-bp, 138-bp, and 141-bp repeats, the ratios were 0.38, 0.51, 0.50, and 0.63, respectively. For the 15-bp repeat, only one nonsynonymous mutation occurred, and d_N/d_S was therefore equal to infinity. We determined that these values were significantly different from 1.0 ($P < 0.001$; *z*test; see Methods for details of analysis). In contrast, the d_N/d_S ratios of the nonrepetitive regions of the ORF were 0.07 (5′ domain), 0.45 (3′ domain upstream of VASES region), and 0.91 (3′ domain downstream of VASES region). Comparing the two sets of ratios, it is apparent that the repeat units roughly matched adjacent nonrepetitive parts of the ORF in terms of d_N/d_S ratio.

Figure 3 Schematic representation of *ALS7* alleles from 67 *C. albicans* strains investigated. The arrangement of allele 11a502 is based on the published sequence (Hoyer and Hecht 2000). The arrangement of all other alleles was deduced from the PCR assays described in Results. Alleles are arranged according to size, using the same symbols as in Figure 1. Names of alleles are in the center of each allele. In naming the allele, the first two digits indicate the number of tandem repeats, followed by an 'a', followed by a single digit indicating the number of stretches of VA/TSES repeats present. In cases of alleles that have the same number of VA/TSES stretches but where the number of the repeats per stretch differs between the alleles, two additional digits are used. The following table gives the number of VA/TSES repeats in each stretch or all alleles shown.

code	n° of VA/TSES repeats in stretches	code	n° of VA/TSES repeats in stretches
a0	0	a304	5,3,2
a101	7	a305	1,19,2
a102	8	a401	2,2,2,2
a103	10	a402	4,2,2,2
a201	2,2	a501	3,2,2,2,2
a202	3,2	a502	5,4,4,3,4
a203	4,3	a6	3,2,2,2,2,2
a301	3,2,2	a7	3,2,2,2,2,2,2
a302	3,5,2	a8	3,2,2,2,2,2,2,2
a303	3,4,2		

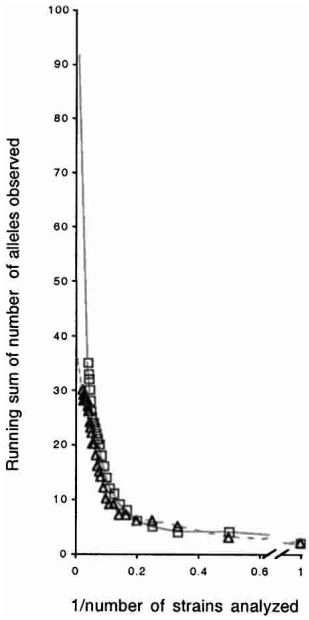


Figure 4 Running sum of the number of different alleles found in cluster strains (triangles, dashed line) and in noncluster strains (squares, solid line) plotted against 1/(running sum) of strains analyzed. The intercept of the trend lines with the *y*-axis gives an indication of the number of alleles that should be observed if an infinite number of isolates were analyzed.

Alleles Present in the Same Strain Are More Similar to Each Other Than Expected by Chance

When we analyzed allele combinations present in individual strains, we discovered that the two alleles in a given strain were significantly more similar (in regard to both tandem repeat domain and VASES region) than would be expected by chance; that is, if all alleles could combine randomly. We detected this by comparing all observed differences between alleles within the same strains with a matrix describing the differences that would be expected if all alleles could be paired randomly, taking allele frequency into consideration (Table 2). When GPG cluster and noncluster strains were considered separately, only the tandem domains in the same strain were significantly more similar to each other than expected by chance (Table 2).

These results differ from the random assortment expected, regardless of the outcome of the current debate on *C. albicans*'s mode of reproduction: clonal and/or (para)sexual. If *C. albicans* still undergoes (para)sexual recombination, alleles should be mixed in the process. If sexual recombination were absent, each of the two present-day alleles in a given strain should be the result of independent divergence of separate clonal lineages, each originating from a single allele in the ancestor of the species (Tibayrenc 1997; Hull et al. 2000; Welch and Meselson 2000).

Two scenarios would explain why alleles in a given strain are more similar to each other than expected by chance. In the first, homozygous *ALS7* alleles, generated by chromosome loss followed by duplication of the remaining chromosome (Janbon et al. 1999) or mitotic recombination leading to duplication of a part of the chromosome (Whelan and Soll 1982), subsequently diverge due to rearrangements and point mutations in the repeat regions. In the second scenario, two different alleles are combined in the same strain as a result of (para)sexual events (either recent or in the past); they are then altered through rearrangements of repetitive regions so that the alleles become more similar (alternatively, possession of highly similar alleles favors [para] sexual events between strains). It should be possible to distinguish between the two scenarios, based on point mutations in the nonrepetitive parts of the gene. Because the frequency of the events leading to homozygosity exceeds the rate of nucleic acid substitution in nonrepetitive DNA (Whelan and Soll 1982; Janbon et al. 1999), under scenario 1 the nonrepetitive parts of two copies of the gene in the same strain should be identical or near identical. At the very least they should be significantly more similar to each other than to nonrepetitive parts of the gene in other closely related strains. Under scenario 2, one would expect to find point mutations distinguishing the nonrepetitive regions of the two alleles in the same strain. We therefore sequenced nonrepetitive parts of the two alleles, 16a201and 17a103, in strain RIHO13. We found eight point mutations, which distinguished nonrepetitive parts of the two alleles. We also found that both alleles were more similar to the nonrepetitive part of the gene in the closely related strain SC5314 than to each other (seven and five substitutions, respectively, separated the two alleles from the SC5314 sequence; both RIHO13 and SC5314 are GPG strains). It thus seems unlikely that alleles 16a201and 17a103 are derived from a recent common ancestral allele.

Detection of *ALS7* mRNA in *C. albicans*

The presence of *ALS7* mRNA in RNA samples from *C. albicans* strains grown in YEPD broth was investigated by two-step reverse transcriptase-polymerase chain reaction (RT–PCR) amplification of the VASES region (Fig. 7A). The RNA samples from all twelve strains tested contained *ALS7* mRNA of the sizes expected based on characterization of the alleles, indicating that none of the various *ALS7* alleles were pseudogenes. Amplification of a *C. albicans EF1B* fragment (spanning an intron) from these RNA samples, and PCR amplification without the reverse transcriptase step, indicated that amplicons were not due to DNA contamination of the samples. *ALS7* mRNA could also be detected in cells grown under a number of other growth conditions. *ALS7*s amplicons were present after one-step RT–PCR amplification of RNA samples from cluster strains AU90 and CLB42 and noncluster strains CLB44 and OD8824 grown on defined (YNB) medium, using primers MC5spF and MC5spR. Cells grown under these conditions were in the yeast morphology. There was no obvious difference between the expression levels in cluster and noncluster strains. *ALS7* mRNA could also be detected by RT–PCR of RNA from the same strains when they were induced to form filaments by first growing them in YNB medium without glucose at pH 6.5 for 3 h and then adding glucose to a final concentration of 8mM (Holmes and Shepherd 1988). There was no obvious difference in *ALS7* mRNA expression between cells grown as yeast or as mycelia, even though hypha-specific expression has been demonstrated for *ALS3* and *ALS8* (Hoyer et al. 1998a; Murad et al. 2001).

ALS7 amplicons were also obtained after RT–PCR amplification of RNA obtained from swabs of oral candidosis lesions taken from five different patients. This indicated that *ALS7* mRNA was expressed in a situation where *C. albicans* is causing human disease. In each case, transcripts corresponding to multiple noncluster alleles were present.

In all of the above experiments, the variable VASES region of the transcript was amplified (using primers MC5spF and MC5spR) in order to confirm that each allele was expressed. Although we were unable to generate a product comprising the entire predicted transcript (even in CLB44 with a short 3a301 allele), it was possible to amplify various portions of *ALS7* mRNA by RT–PCR, including the N-terminal portion of the gene (nt 401–1319 of the 6006-nt ORF, using primers MC5N2F and MC5NR). This indicated that mRNA of the expected size was present in *C. albicans* strains. We directly confirmed the presence

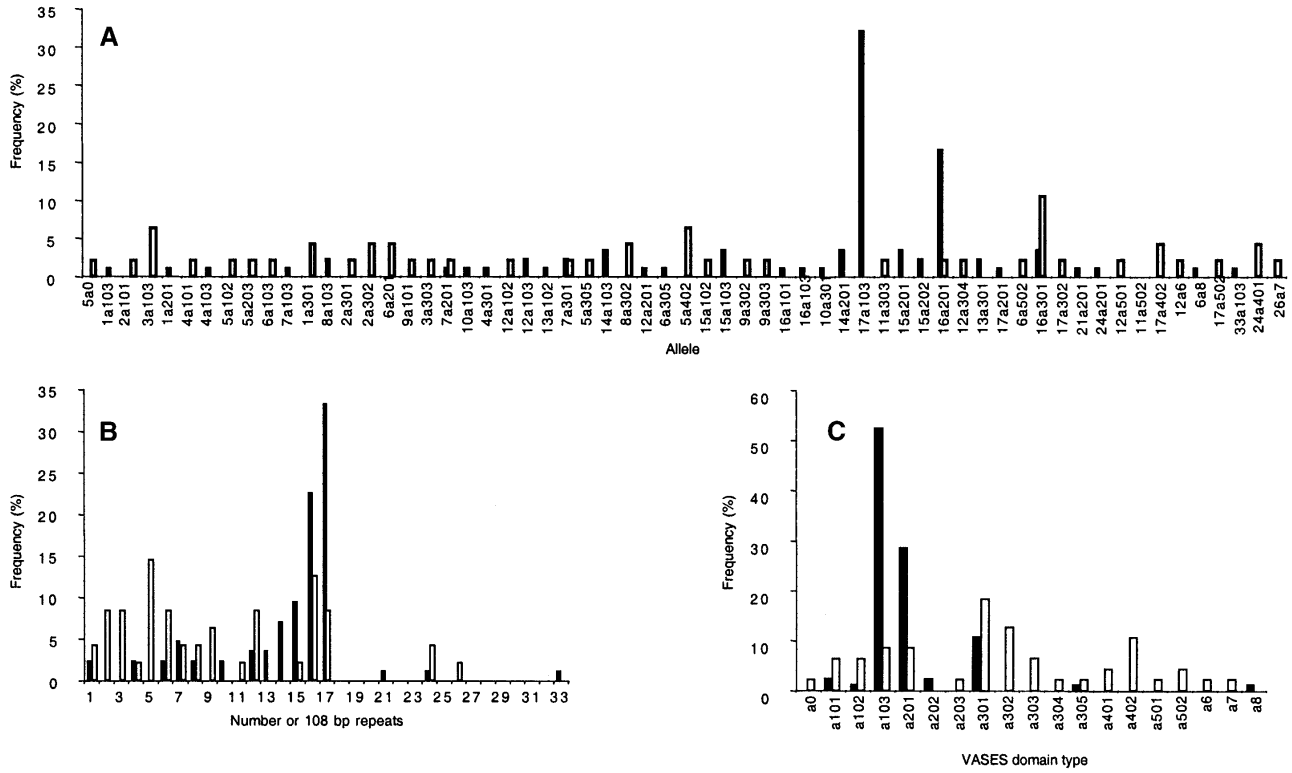


Figure 5 (*A*) Frequency of alleles, (*B*) frequency of different sized tandem repeat domains, and (*C*) frequency of different types of VASES regions for GPG cluster strains (solid bars) and noncluster strains (hollow bars) analyzed. Allele frequencies were calculated assuming that all strains tested were diploid; that is, where PCR analysis suggested the presence of only one allele in a strain, the allele was assumed to be present in two copies. The two borderline strains (HUN92 and HUN122; see a list of strains in Suppl. Material) were included in the calculations.

of full-length *ALS7* mRNA in two strains, SC5314 and Hun68, using Northern blots and probes that hybridized with either the VASES region or the tandem repeat domain. Both probes hybridized to RNA of a size expected for the full-length transcript. However, transcript concentrations were low in both strains under three different growth conditions tested (YPD medium, RPMI medium, and YPD medium with bovine serum added; the latter two conditions induce *ALS3* expression [Hoyer et al. 1998a]). Based on comparisons of the *ALS7* hybridization signal intensity with that of actin mRNA (see Fig. 7B for example), taking into account effectiveness of transfer of mRNA from gel to filter, and the size of hybridizing region of the mRNA molecules, we estimate that the concentration of *ALS7* mRNA was approximately 1%–2% that of actin mRNA. This is comparable to the basal (uninduced) mRNA concentration of an *S. cerevisiae* gene with a comparable function, the *AGA1* gene encoding the anchorage subunit of agglutinin (Roy et al. 1991; http://genome-www4.stanford.edu/cgi-bin/SGD/SAGE/querySAGE).

We note that detection of transcript by Northern hybridization required the use of long probes and the loading of large amounts of mRNA (see Methods). Initial attempts to detect *ALS7* mRNA in a variety of strains by Northern hybridization using shorter probes and loading less RNA failed, although *ALS3* message was detected (data not shown). This indicates that low levels of transcription of *ALS7* in cultures were not restricted to SC5314 and Hun68.

Expressed Reverse ORFs Within *ALS7*

Analysis of available nucleotide sequences of *ALS7* genes (SC5314, 1161, two alleles of RIHO13, and partial sequence of var1.7) revealed that each contained ORFs, initiated with an ATG, in the opposite orientation to *ALS7*. These were located almost entirely within the tandem repeat domain. SC5314 had one reverse ORF of 517 amino acids (*ALS7* nt 2807–1254; see Fig. 1). The two alleles for GPG cluster strain RIHO13 each contained two adjacent reverse ORFs that were in-frame and separated by a stop codon and 6 bp. The lengths of the ORFs were 429 and 145 aa in allele 17a103, and 333 and 217 aa in allele 16a1201. Noncluster strain var1.7 had one reverse ORF of 157 aa, and noncluster strain 1161 had two adjacent, in-frame, reverse ORFs of 153 and 73 aa. An amino acid sequence alignment of all putative proteins encoded by the reverse ORFs showed that they had a high degree of homology. Excluding deletions, the average number of amino acid substitutions between two sequences was 19 ± 14, and the maximum was 44.

To determine whether the reverse ORFs could encode functional proteins, we measured d_N/d_S ratios and investigated their amino acid sequence similarity to other proteins. The d_N/d_S ratio among the five reverse ORF sequences was 0.67, significantly different from 1.0 ($P < 0.001$, binominal test; see Methods for analysis details). In a BLAST search against nonredundant GenBank protein sequences, the best match between the reverse ORFs was in most cases *C. albicans* Als7p. The exceptions were the reverse ORFs of var1.7 (Als7p was the second best match), the second reverse ORFs in both RIHO13 alleles (no match with Als7p but with other proteins), and the second ORF in 1161 (no significant match with any protein in the database).

Further analysis suggested that the reverse ORFs could be transcribed. In SC5314 there was a TATA box positioned 106 nt upstream of the reverse ORF, and two CAAT boxes −118 and −146 relative to the start codon. Significantly, both the TATA

Table 1. Nucleotide Substitution Rates in Various Regions and Units of the *ALS7* Gene

Region/unit	Average ± SD nucleotide substitutions per bp between units[a] (no. of nucleotides analyzed)	Strains analyzed
5′ nonrepetitive domain	0.002 ± 0.001 (4 × 1311 = 5244)	SC5314, 1161, RIHO13 (two alleles)
3′ domain, nonrepetitive region upstream of VASES region	0.007 ± 0.007 (5 × 900 = 4500)	SC5314, 1161, var1.7 RIHO13 (two alleles)
3′ domain, nonrepetitive region downstream of VASES region	0.003 ± 0.001 (5 × 1389 = 6945)	SC5314, 1161, var1.7, RIHO13 (two alleles)
108 bp tandem repeat (all)	0.156 ± 0.104 (64 × 108 = 6912)	SC5314, 1161, var1.7, RIHO13 (two alleles)
First 108 repeat	0.004 ± 0.005 (5 × 108 = 540)	SC5314, 1161, var1.7, RIHO13 (two alleles)
Last 108 repeat	0.019 ± 0.024 (5 × 108 = 540)	SC5314, 1161, var1.7, RIHO13 (two alleles)
Other 108 repeats	0.101 ± 0.055 (54 × 108 = 5832)	SC5314, 1161, var1.7, RIHO13 (two alleles)
All VA/TSES repeats	0.034 ± 0.033 (83 × 15 = 1245)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c
First pair of VA/TSES unit stretch[b]	0.000 ± 0.000 (23 × 30 = 690)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2C
VASES repeats	0.000 ± 0.000 (41 × 15 = 615)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, OD8824, W132
VTSES repeats	0.000 ± 0.000 (42 × 15 = 630)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, OD8824, W132
87 bp repeat (all)	0.025 ± 0.015 (43 × 87 = 2697)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, gaymc-c, var1.7
87 bp repeat (first)	0.007 ± 0.017 (13 × 87 = 1131)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, gaymc-c
87 bp repeat (middle)	0.022 ± 0.014 (16 × 87 = 1392)	SC5314, 1161, RIHO13 (one allele), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, gaymc-c
87 bp repeat (last)	0.008 ± 0.008 (14 × 87 = 1213)	SC5314, 1161, RIHO13 (two alleles), CLB44, FJ26, jam2c, gaymc-c, var1.7
138 bp repeat	0.019 ± 0.013 (28 × 138 = 3864)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, OD8824, W132
141 bp repeat (all)	0.053 ± 0.047 (29 × 141 = 4089)	SC5314, 1161, RIHO13 (two alleles) CLB42, CLB44, Au90, FJ9, FJ26, jam2c, gaymc-c, OD8824, W132
141 bp repeat (last)	0.023 ± 0.045 (12 × 141 = 1692)	SC5314, 1161, RIHO13 (two alleles), CLB42, CLB44, Au90, FJ9, FJ26, jam2c, gaymc-c, OD8824, W132

[a]Data from "uncorrected p" distance matrices generated by Paup*.
[b]Stretch is an uninterrupted series of VA/TSES units.

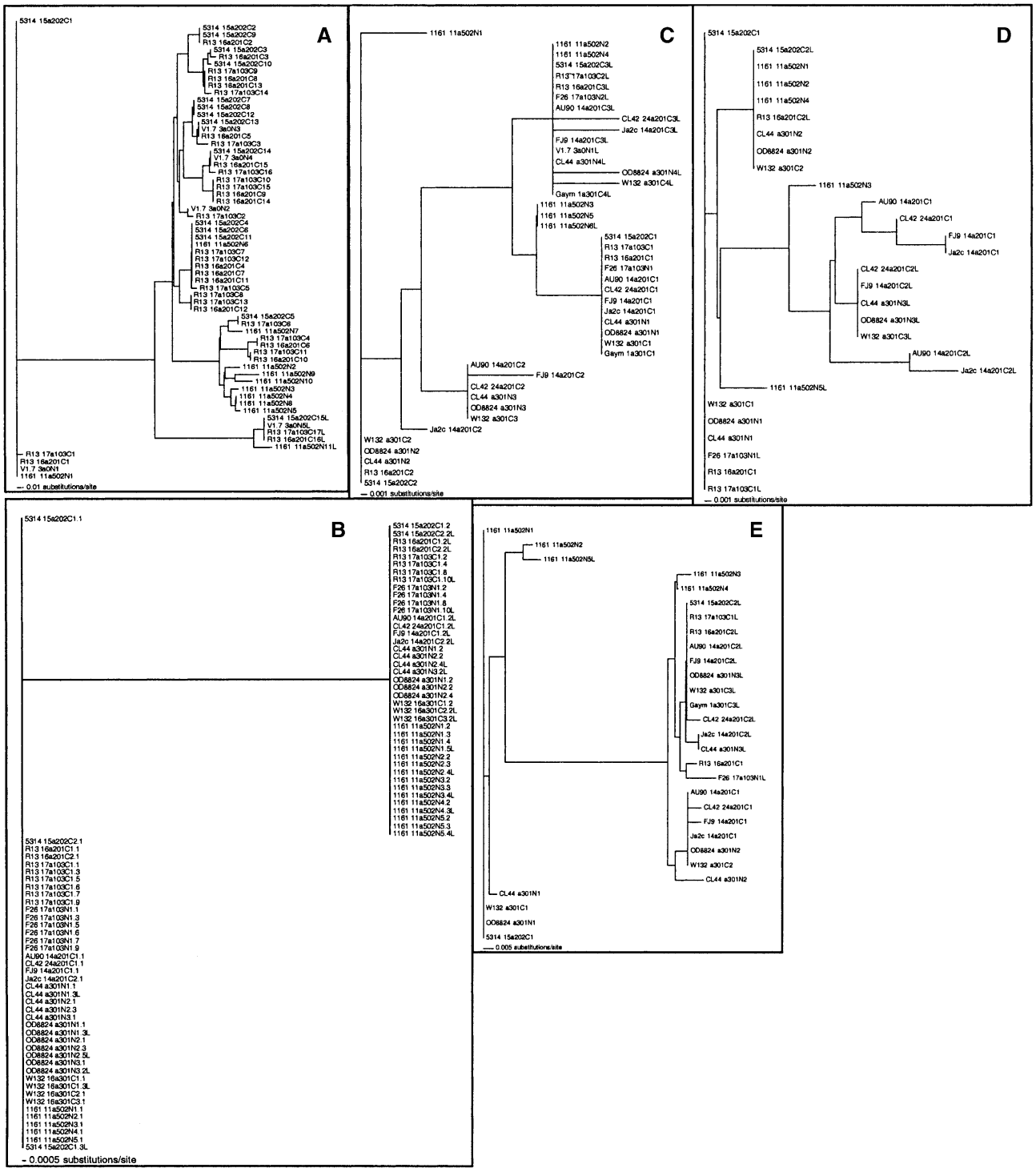


Figure 6 Neighbor-joining trees based on genetic distances (uncorrected p) between sequences of individual 108-bp repeats (*A*), 15-bp VA/TSES repeats (*B*), 87-bp repeats (*C*), 138-bp repeats (*D*), and 141-bp repeats (*E*). Repeats are labeled as follows: The first digits indicate the strain (up to four digits; names are abbreviated if necessary), followed by a blank space, followed by the name of the allele (using the same nomenclature as in Fig. 3; in some cases the name is incomplete because only one VASES region was cloned and sequenced, and the allele to which it belonged was not determined), followed by a C if the strain belongs to the GPG cluster, and an N for noncluster. In all panels except *B*, the next digits give the position of the repeat in the domain, that is, the first unit of its kind in the domain has the number 1. The last unit of its kind in a domain is labeled with an L as the final letter. In *B*, the position of VA/TSES repeats is designated by two numbers: The first describes the number of the stretch of VA/TSES units in the region, and the second describes the position of the repeat in the stretch; this is followed by an L if it is the last unit in the stretch (e.g., 2.3L is the third and last unit in the second stretch of VA/TSES repeats in the region).

box and one of the CAAT boxes lie in the last 108-bp repeat of *ALS7*, which is distinct from other 108-bp repeats (see above), and only the last 108-bp repeat has these features. There was also a motif, TAG…TATGT…TTT, 44–46, 53–57, and 102–104 nt, respectively, downstream of the ORF stop codon, that matched the consensus for *S. cerevisiae* transcription termination sequences (Zaret and Sherman 1982). The sequence TATAAA, 193 nt downstream of the tripartite transcription termination signal, could possibly function as a polyadenylation motif. Similar promoter, termination, and polyadenylation motifs were also present in both RIHO13 alleles and var1.7 (in the latter strain we did not sequence the region where the termination and polyadenylation signals are predicted to lie, but they are likely to be present because they are located in a nonrepetitive, highly conserved part of the gene). Noncluster strain 1161, whose last 108-bp repeat differs from that found in the other strains (see above) had only the transcription termination signals downstream of the combined ORFs, but no TATA or CAAT box upstream of the ORFs.

We investigated whether mRNAs corresponding to the reverse ORFs were indeed present in cells, using four GPG cluster strains (SC5314, RIH013, Jam-2c, OD8807) and four noncluster strains (Var1.7, W53, gaymc-c, YsU123). RT–PCR was carried out as a two-step process. cDNA was synthesized from RNA in two reactions, one using primer TRF which bound to, and copied, *ALS7* mRNA, the other using TRR which bound to, and copied, the reverse ORF. As the predicted 5′ end of the reverse ORF is within the tandem repeat domain, primer TRF bound several times within this domain. TRF, however, was specific for *ALS7* mRNA and would only bind to the reverse ORF mRNA with 7/23 mismatches. Likewise, TRR was specific for the reverse ORF mRNA and would only bind to *ALS7* mRNA with 11/24 mismatches. The cDNAs were then PCR-amplified using TRF and TRR as primers. There was a ladder of amplicons in both the *ALS7*-specific and the reverse ORF-specific cDNAs in all of the strains (results for four strains are shown in Fig. 8), indicating that both *ALS7* and the reverse ORF mRNA were expressed in these cells. The distance between the ladder bands was 108 bp, which is equivalent to the repeat length in the tandem repeat domain. The lack of the two smallest reverse ORF amplicons from SC5314 was due to point mutations in two of the tandem repeats in this strain that prevented the TRF oligonucleotide binding and priming amplification.

We were not able to detect the reverse ORF mRNA on Northern blots. These blots were probed with DNA probes corresponding to the entire tandem repeat region (1.6 kb and 2.6 kb respectively in strains Hun68 and SC5314), which should hybridize to both forward and reverse message. Taking into account the strength of the signal for the *ALS7* message at various amounts of mRNA loaded, the concentration of reverse ORF transcript must be less than15% of that of the forward transcript.

DISCUSSION

Evidence That *ALS7* Is a Hypermutable Contingency Locus

Highly variable surface proteins are found in many microbes, in particular in pathogens such as *Plasmodium*, *Giardia*, and *Trypanosoma* (Duncan et al. 1991; Ramasamy 1998; Ayala and Rich 2000; Adam 2001). The genes encoding these proteins have been termed 'hypermutable contingency genes', and the major mechanism underlying the protein variability is rearrangement of repeat-containing parts of their ORFs (Duncan et al. 1991; Moxon and Thaler 1997; Barry and McCulloch 2001; Bridges 2001). *ALS7* meets the criteria for hypermutable contingency loci in that (1) it is part of a family of surface proteins (Hoyer 2001), (2) it contains repeat units which make it potentially highly mutable, and (3) rearrangement of conserved units in repeat regions generates allelic diversity far in excess of the diversity in both nonrepetitive parts of the *ALS7* ORF and far in excess of the diversity reported for other *C. albicans* ORFs (Bougnoux et al. 2002).

Biological Functions of *ALS7* and *ALS7* Variability

Hypermutable contingency loci generate phenotypic variants with a frequency of approximately 10^{-3} switch/organism/population doubling—the frequency is fairly constant among different microbes of both eukaryotic and prokaryotic origin (Barry and McCulloch 2001). The resulting repertoire of phenotypes is believed to assist in the survival of the organism by allowing it to rapidly respond to, and exploit, alterations in unpredictable and/or hostile environments (Duncan et al. 1991; Moxon and Thaler 1997; Barry and McCulloch 2001; Bridges 2001). In the pathogens *Giardia*, *Plasmodium*, and *Trypanosoma*, the biological role of hypermutable surface proteins has been the subject of considerable study. It is generally agreed that evasion of the immune system is one function of these proteins. This is achieved by escaping an immune response through alteration of immunogenic epitopes and also possibly by interfering with maturation of antibody responses, because of cross-reactivity between different forms of repeat domains (Duncan et al. 1991; Nash 1997; Ramasamy 1998; Adam 2001). Ramasamy (1998) also discusses a possible function as superantigens. However, it is also widely believed that immune evasion is not the only function of these proteins (Nash 1997; Ramasamy 1998;

Table 2. Observed Differences Between Two Copies of the *ALS7* Locus in the Same Strain and Differences Expected from Random Mixing of Alleles

	Differences[a] observed[b]; average ± SD (n)	Differences[a] expected[c] average ± SD (n)	Significance (*P*)[d]
Tandem repeat domain			
All	2.0 ± 3.1 (66)	6.4 ± 5.4 (17424)	8×10^{-11}
GPG cluster strains	1.5 ± 1.9 (42)	4.4 ± 4.9 (7056)	0.0001
Noncluster strains	2.8 ± 4.3 (24)	7.2 ± 5.7 (2304)	0.0002
VASES region			
All	282 ± 352 (66)	447 ± 464 (17424)	7×10^{-15}
GPG cluster strains	231 ± 297 (42)	263 ± 394 (7056)	0.5915
Noncluster strains	373 ± 424 (24)	553 ± 466 (2304)	0.0588

[a]Differences are expressed as differences in the numbers of repeats for the tandem repeat domain and as differences in the sizes (bp) of the VASES domain.
[b]Average differences between alleles in the same strain as observed for all strains analyzed in this study. It is assumed that all strains were diploid and that a strain was homozygous if only one allele was observed.
[c]A matrix was generated, describing pairwise differences between all possible combinations of the alleles observed in the strains. Each allele was used for comparisons as often as it occurred, so that the size of the matrix was equal to the square of 2 × no. of strains analyzed; thus the contribution of an individual allele to the average of differences was dependent on the frequency with which it was observed. It is assumed that all strains were diploid and that a strain was homozygous if only one allele was observed.
[d]Two-sided *t*-test comparing observed differences and expected differences.

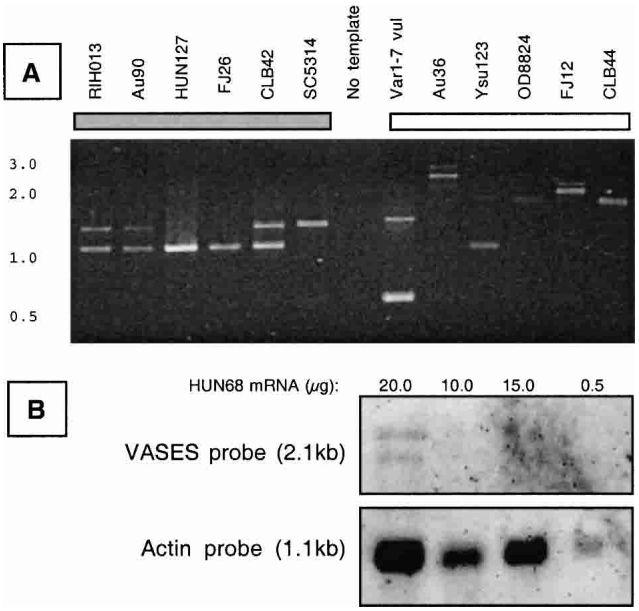


Figure 7 Expression of *ALS7*. (*A*) RT–PCR amplification of VASES region of *ALS7* mRNA from GPG cluster and noncluster *C. albicans* strains. Cells were grown in YEPD broth, and RT–PCR used primers MC5spF and MC5allF. Numbers on the *left* indicate DNA fragment size in kb. (*B*) Northern hybridization verifying full-size transcripts (7.2 kb and 7.8 kb) in Hun68 cells grown in YPD plus bovine serum. Varying amounts of mRNA were loaded, and blots were hybridized with a 2.1-kb probe generated from the HUN68 VASES region (the same bands were seen when the tandem repeat domain was used as a probe; data not shown), stripped, and reprobed with a 1.1-kb actin probe. Exposure time was 60 min.

Adam 2001). Nash (1997) suggests that the nature of the genes encoding these proteins allows the rapid generation and selection of sets of clonal lines which are best suited to adhere to different host surfaces. Nash (1997) also points out that highly variable surface proteins with repeat domains are not restricted to pathogens but can be found in free-living amoeba, supporting a role in adhesion, or other functions not related to immune selection. There is also evidence of an important role of repeat-containing surface proteins in adhesion of bacteria when colonizing human hosts (McNab et al. 1996, 1999; Roos and Jonsson 2002).

Our analysis of *ALS7* indicates a main role in adhesion rather than in immune evasion for several reasons. First, expression levels seem to be low under many growth conditions, and we would expect an 'immunological decoy' to be expressed at higher rates. Second, the constraints restricting the arrangements of most of the repeat units would suggest a role beyond that of a decoy. Third, if the diversity-generating repeats had mainly a decoy function, one might expect them to have higher d_N/d_S ratios than the remainder of the ORF, because amino acid substitutions in them would increase diversity and could be expected to have a high probability of enhancing fitness. This is not what we observed. Rather, d_N/d_S ratios of repeat units roughly matched those of adjacent nonrepetitive parts of the ORF, suggesting that the parts of Als7p encoded by the repeats are under similar functional restraints as the parts encoded by adjacent nonrepetitive units. A role of Als7p in adhesion would also be in keeping with both functional and expression data on other Als proteins (Gaur and Klotz 1997; Chandra et al. 2001; Hoyer 2001; Fu et al. 2002).

Similar arguments regarding biological role may apply to the putative additional proteins, encoded by the reverse ORFs within the gene. However, we have no direct evidence that the reverse ORFs are translated. Only the reverse ORFs' low d_N/d_S ratio, and the significant similarity between their predicted products and Als7p, suggest that the reverse ORFs could encode proteins, and that their role may be related to that of the Als proteins.

A case can be made that maintenance of particular arrangements of repeat units in the *ALS7* gene is driven by continuous selection in the human host, and that therefore the role of the gene and its product(s) is specific to the interaction between *C. albicans* and the human host. This argument is based on differences at the *ALS7* locus between the two laboratory strains 1161 and SC5314 on the one hand and the remaining isolates, none of which were propagated for prolonged periods outside the host (our patient isolates are stored at −70°C and were not subcultured extensively before analysis), on the other. The *ALS7* allele of strain 1161 often violates restraints on the arrangements of repeat units common to other strains (as illustrated in Fig. 6) and is missing transcription initiation signals for the reverse ORFs. Likewise SC5314, although a GPG cluster strain, has a VASES region that does not produce the GPG-specific AFLP product present in 90% of GPG patient isolates. It is also unusual in that it only has a single reverse ORF. A simple explanation for these findings is that the two laboratory strains, while being subcultured in various laboratories, underwent modifications in hypermutable sequence regions that would have been selected against in the host. If so, it should be possible to generate, in the laboratory, clonal lineages of a given strain with different *ALS7* alleles and use competition experiments between these lineages in an animal model to learn more about the biological function of the repeat units and their arrangements.

Transcriptional Control of *ALS7* Expression?

Expression of *ALS7* mRNA could be detected in *C. albicans* cells cultured under a variety of conditions and in cells obtained directly from oral candidosis lesions. It would be interesting to see whether *ALS7* mRNA can be detected in oral samples taken from people colonized with *C. albicans* without clinical signs of candidosis (as has been shown for some of the secreted aspartic proteinase genes; Naglik et al. 1999). The level of *ALS7* expression in vitro was low and comparable to that of a gene encoding a related yeast cell surface protein under uninduced conditions. It seems likely, therefore, that under our in vitro experimental conditions, *ALS7* transcription was repressed. Although we were able to detect the reverse ORF mRNA by RT–PCR, we were unable to detect the mRNA directly on a Northern blot. This indicates that the reverse ORF mRNA is present in cells, but at lower concentrations than *ALS7* mRNA. The possibility that it may play a role in the control of *ALS7* expression (Vanhee-Brossollet and Vaquero 1998) requires further investigation.

If the function of Als7p is in adhesion, why is its expression regulated to such a low level under a variety of growth conditions? A likely explanation is that Als7p has a highly specialized function. *C. albicans*, like many other microorganisms, possesses several putative adhesins that enable it to adhere to, and colonize, a variety of surfaces (Cannon and Chaffin 1999). Many of these adhesins are large-molecular-mass glycoproteins, the production of which puts a metabolic burden on the cells. Constitutive Als7p expression may constitute a higher cost to the cell than mechanisms regulating its synthesis. An additional benefit of such regulation is that it will minimize the development of neutralizing antibodies by the host. The idea of a specialized function for *ALS7* would make sense, given that it is part of a family of related adhesins and given that several other *ALS* genes are known to be transcriptionally activated by different sets of stimuli (Hoyer et al. 1998a,b; Chandra et al. 2001; Murad et al. 2001).

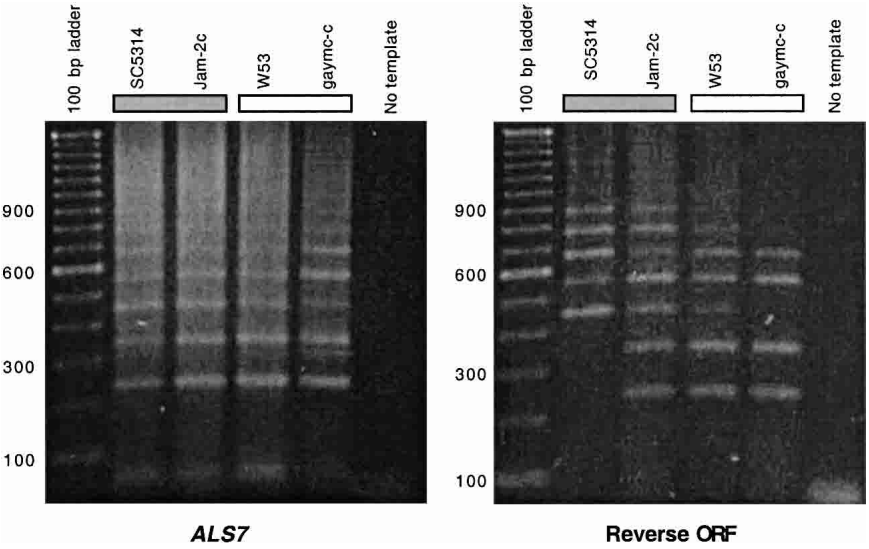


Figure 8 RT–PCR amplification of *ALS7* and reverse ORF mRNA from *C. albicans* strains. *ALS7* cDNA was synthesized from RNA using primer TRF and reverse ORF cDNA was synthesized from RNA using primer TRR. cDNA was amplified by PCR using primers TRF and TRR. Numbers on the *left* indicate DNA fragment size in base pairs.

Evidence That Als7p Contributes to the Success of GPG Cluster Strains

We began this study with the aim of identifying genes that contribute to *C. albicans* pathogenicity, by studying polymorphisms specific to GPG cluster strains—which cause infections significantly more often than other *C. albicans* genotypes. The low genetic diversity among GPG cluster strains suggests that these strains could be of fairly recent origin (Schmid et al. 1999). Therefore it is possible that some GPG cluster-specific polymorphisms are not related to pathogenicity but are present merely because all of these strains share a recent ancestor. GPG cluster-specific features of the *ALS7* gene, however, are unlikely to be a result of derivation from a recent common ancestor. Given that cluster strains have the same set of repeat elements as noncluster strains, and given the high rates of intragenic recombination in repeat regions (10^{-2} to 10^{-3} per generation; Levinson and Gutman 1987; Schug et al. 1998; Ayala and Rich 2000), it is unlikely that ancestral features of the gene would have been maintained in cluster strains, unless they were continually selected for (this is also supported by the analysis of the laboratory strains, especially 1161); that is, unless they and the *ALS7* gene that contain them are contributing to the success of these strains as pathogens.

Related to this we would also note that, although we were able to demonstrate genetic differences between GPG cluster strains from different geographic regions using the Ca3 probe (Schmid et al. 1999), we found no indication that isolates from different geographic regions differ in regard to their *ALS7* genes (data not shown). Unless cluster-specific features of the *ALS7* genes are under constant selection, it is difficult to explain why geographically separated and genetically distinct lineages of GPG cluster strains are similar in regard to a hypermutable sequence.

If particular features of the *ALS7* gene product(s) contribute to the success of cluster strains, they probably do not do so in isolation. There is no apparent reason why noncluster strains could not generate alleles at least very similar to those present in cluster strains. If so, the low prevalence of such alleles in noncluster strains must mean that their possession does not confer a selective advantage on noncluster strains, but only on cluster strains. Therefore there must be cluster- specific polymorphisms in other genes, which are required for the *ALS7* polymorphisms to convey a selective advantage. Polymorphisms in other *ALS* genes would be obvious candidates. Sequence and functional analyses of allele repertoires of the different *ALS* genes in individual patient isolates and functional analysis of strains with experimentally altered sets of allelic repertoires could be a valuable approach to elucidate the function of this gene family.

METHODS

Strains

Strains used are listed in Supplemental Material (available online at http://www.genome.org). Two overlapping sets of patient isolates (25 GPG cluster strains and 16 noncluster strains) were selected to represent a collection of 266 infection-causing isolates from 12 geographic regions in six countries, genotyped using the probe Ca3 (Schmid et al. 1999). Strains for set one were selected using the quartet method (Schmid et al. 1999) as follows: The Ca3-based genetic distances were calculated for all possible quartets containing two GPG cluster strains and two noncluster strains from the collection. For each GPG cluster strain, the number of times it was closest to another cluster strain in these quartets was recorded. Likewise, for each noncluster strain the number of times it was closest to the other noncluster strain in these quartets was also recorded. The higher the quartet value, the more distinct the GPG cluster strain was from noncluster strains and the more distinct the noncluster strain was from GPG cluster strains. After strains were thus ranked by quartet values, high-ranking GPG cluster and noncluster strains from each geographic region were chosen as model strains. Set 2 was chosen on the principle that the genetic distances (based on Ca3 fingerprints) between all GPG cluster strains and the 25 model GPG cluster strains were as small as possible, and the genetic distances between all noncluster strains and the 16 noncluster model strains were also as small as possible, using algorithms developed by Holland (2001). Because different geographical regions were represented in the collection by different numbers of isolates, a correction was made to ensure that each region had the same impact in the process. This was achieved by weighting the contribution of each strain to the distance calculations so that it was inversely proportional to the number of isolates from the region. Laboratory strain SC5314 was added to the sets as the 26th GPG cluster strain (based on Ca3 fingerprinting data; Giblin et al. 2001).

General Molecular Biology Methods

Methods were generally based on those described by Ausubel et al. (1994). All PCR reactions were performed in a final volume of 20 µL containing 1 U *Taq* DNA polymerase (QIAGEN), 4 µL of Q-buffer, and 1× PCR buffer supplied by the manufacturer (QIAGEN), 10 pmol of each primer, 200 µM of each dNTP (Roche Diagnostics), and 10–100 ng DNA. PCR conditions were varied according to the primers used (a list of primers can be found in the Suppl. Material), following the guidelines listed in Ausubel et al. (1994). All PCR protocols included a final 5-min extension step at 72°C. In PCR amplifications using primers MC5pf and MC5outpr, a 'touchdown' protocol was used (Don et al. 1991); that is, the annealing temperature was 10°C above the optimum annealing temperature for cycles 1–6 and 5°C above the optimum for cycles 7–12 (the total number of cycles was 36).

DNA Extraction, AFLP Analysis, and Isolation of the MC5-c AFLP Product

C. albicans DNA was extracted from strains using the method of Scherer and Stevens (1987) and was used as a template for AFLP as described by Vos et al. (1995). The AFLP used *Mse*I adapters only, and primers M-C (preselective amplification) and M-CTC (see list of primers in Suppl. Material). AFLP products were separated by electrophoresis in 5% (w/v) denaturing polyacrylamide gels (Maxam and Gilbert 1980) and visualized by silver staining (Promega Silver Staining System, Promega). The GPG cluster-specific band MC5-c was excised from the gel, re-amplified using Primer M-CTC, gel purified, cloned into pGEM-T® (Promega), and sequenced, using universal primers M13F and M13R, on an ABI Prism 377 sequencer (Perkin Elmer).

Characterization of VASES Regions

Initially the entire region was amplified using primers MC5pf and MC5outpr, and its size assessed on 0.8% agarose gels, using 1Kb Plus DNA Ladder (Invitrogen) and previously characterized VASES regions of known size as molecular weight standards. The number of units of 366 bp (87-bp repeat + 138-bp repeat + 141-bp repeat) that would equate to this size, taking into consideration that at the beginning of the region there will be an 87 + 138-bp unit and at the end a 141 + 87-bp unit was calculated (see Fig. 1 for general organization of VASES regions). The PCR product was then used as a template for PCR amplifications with primers MC5VApf and MC5VApr, and primers MC5spF and MC5Vapf, to confirm the proposed structure and to determine the size of the various stretches of 15-bp VA/TSES units in that region.

Calculation of d_N/d_S Ratios

The software PAML (Yang 1997) was used to determine the ratio of synonymous/nonsynonymous changes (d_N/d_S) between pairs of sequences (Yang and Nielsen 2000). PAML can only calculate ratios for sequences without deletions, and in case of the reverse ORF only a 135-bp deletion-free region present in all ORFs was therefore used for analysis. Because two identical sequences or two sequences differing only by nonsynonymous mutations will generate an undefined d_N/d_S value (division by zero), such values cannot be combined with others to generate a meaningful average; all averages reported in the text therefore exclude such pairs. In order to include as many values as possible in statistical tests aimed at determining whether the average d_N/d_S for a sequence was $\neq 1.0$, the following approach was used: For each pair of sequences compared, the resulting d_N/d_S was categorized as <1, =1, or >1 (cases in which there were only nonsynonymous changes can therefore be included as >1 in the analysis, and only cases where no mutation distinguishes the two sequences are excluded). We then determined whether the frequency of values >1 was significantly different from the frequency of values <1, using the *z* test or binominal test. A significant difference indicates that the values were not symmetrically distributed around a value of 1.0, and that therefore the average ratio observed was significantly different from 1.0.

RT–PCR and Northern Hybridization

RNA was extracted from *C. albicans* cells grown in media as described in the Results section or from *C. albicans* cells isolated directly from oral candidosis lesions. The lesions of patients presenting with oral candidosis at University of Otago School of Dentistry clinics were wiped with a sterile swab. All patients were HIV-negative. Ethical approval for the procedure was obtained from the Ministry of Health Otago Ethics Committee (protocol no. 99/11/095). The swab was immediately vortex mixed in 2 mL YEPD broth, and a portion of the liquid was diluted and plated on YEPD agar containing chloramphenicol (20 µg/mL) to confirm and quantify *C. albicans* presence. The yeasts in the remaining YEPD broth were harvested by centrifugation (3,000 *g*, 5 min) prior to RNA extraction. Total RNA was isolated from *C. albicans* cells (1.8×10^9 cultured cells or all remaining cells isolated from oral swab) by the hot phenol method of Schmitt et al. (1990). For RT–PCR, RNA samples were treated with DNAase I (DNA-free kit, Ambion) to remove traces of DNA from the RNA samples. Purity of RNA samples was confirmed by RT–PCR amplification using primers EF1BF and EF1BR that spanned the intron in *C. albicans EF1B*. DNA present in the sample (containing the intron) gave an amplicon of 891 bp; RNA gave an amplicon of 526. RT–PCR was carried out as either a one-step reaction (OneStep RT–PCR kit, QIAGEN) or a two-step process. The first step in the two-step process was reverse transcription of RNA samples using SuperScript reverse transcriptase (Invitrogen) and a specific primer or a poly-T primer. The RNA was removed from the cDNA with RNase H, and the cDNA was amplified using Taq polymerase and specific primers. For Northern hybridization, serial dilutions of mRNA (0.5–20µg) obtained from total RNA using the Sigma GenElute mRNA isolation kit were loaded on 1%agarose gels, containing 6% formaldehyde. RNA was transferred to nylon membranes (Roche). Each blot was hybridized first with a probe comprising the VASES region, then with a probe comprising the actin gene, and finally with a probe comprising the tandem repeat domain. Double-stranded DNA probes for each Northern blot were prepared from DNA from the corresponding strain by DIG labeling with random priming (the double-stranded tandem repeat domain probe would therefore hybridize with both forward and reverse transcripts). Detection was via luminescence (DIG Nonradioactive Nucleic Acid Labeling and Detection System; Roche). A lane of DIG-labeled RNA size markers (1.4 kb–6.9 kb) was included on every blot.

ACKNOWLEDGMENTS

This work was funded by a Marsden grant from the Royal Society of New Zealand. Sequence data for *C. albicans* SC5314 was obtained from the Stanford Genome Technology Center website at http://www-sequence.stanford.edu/group/candida. This sequencing was accomplished with the support of the NIDCR and the Burroughs Wellcome Fund. We thank Mr. N. Firth (Department of Stomatology, University of Otago) for obtaining the oral candidosis samples described in this paper, and Lois Hoyer and John Tweedie for helpful comments on the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adam, R.D. 2001. Biology of *Giardia lamblia*. *Clin. Microbiol. Rev.* **14:** 447–475.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Seidman, J.G., and Struhl, K. 1994. *Current protocols in molecular biology*. J. Wiley, New York.

Ayala, F.J. and Rich, S.M. 2000. Genetic variation and the recent worldwide expansion of *Plasmodium falciparum*. *Gene* **261:** 161–170.

Barry, J.D. and McCulloch, R. 2001. Antigenic variation in trypanosomes: Enhanced phenotypic variation in a eukaryotic parasite. *Adv. Parasitol.* **49:** 1–70.

Botterel, F., Desterke, C., Costa, C., and Bretagne, S. 2001. Analysis of microsatellite markers of *Candida albicans* used for rapid typing. *J. Clin. Microbiol.* **39:** 4076–4081.

Bougnoux, M.E., Morand, S., and d'Enfert, C. 2002. Usefulness of multilocus sequence typing for characterization of clinical isolates of *Candida albicans*. *J. Clin. Microbiol.* **40:** 1290–1297.

Bridges, B.A. 2001. Hypermutation in bacteria and other cellular systems. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **356:** 29–39.

Cannon, R.D. and Chaffin, W.L. 1999. Oral colonization by *Candida albicans*. *Crit. Rev. Oral Biol. Med.* **10:** 359–383.

Chandra, J., Kuhn, D.M., Mukherjee, P.K., Hoyer, L.L., McCormick, T., and Ghannoum, M.A. 2001. Biofilm formation by the fungal pathogen *Candida albicans*: Development, architecture, and drug resistance. *J. Bacteriol.* **183:** 5385–5394.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., and Mattick, J.S. 1991. "Touchdown" PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* **19:** 4008.

Duncan, L.R., Gay, L.S., and Donelson, J.E. 1991. African trypanosomes express an immunogenic protein with a repeating epitope of 24 amino acids. *Mol. Biochem. Parasitol.* **48:** 11–16.

Fu, Y., Ibrahim, A.S., Sheppard, D.C., Chen, Y.C., French, S.W., Cutler,

J.E., Filler, S.G., and Edwards Jr., J.E. 2002. *Candida albicans* Als1p: An adhesin that is a downstream effector of the EFG1 filamentation pathway. *Mol. Microbiol.* **44:** 61–72.
Gaur, N.K. and Klotz, S.A. 1997. Expression, cloning, and characterization of a *Candida albicans* gene, *ALA1*, that confers adherence properties upon *Saccharomyces cerevisiae* for extracellular matrix proteins. *Infect. Immun.* **65:** 5289–5294.
Giblin, L., Edelmann, A., Zhang, N., Maltzahn, N.B.v., Cleland, S.B., Sullivan, P.A., and Schmid, J. 2001. A DNA polymorphism specific to *Candida albicans* strains exceptionally successful as human pathogens. *Gene* **272:** 157–164.
Holland, B. 2001. "Evolutionary analysis of large data sets: Trees and beyond." Ph.D thesis, Massey University, Palmerston North, New Zealand.
Holmes, A.R. and Shepherd, M.G. 1988. Nutritional factors determine germ tube formation in *Candida albicans*. *J. Med. Vet. Mycol.* **26:** 127–131.
Hoyer, L. 2001. The *ALS* gene family of *Candida albicans*. *Trends Microbiol.* **9:** 176–180.
Hoyer, L.L. and Hecht, J.E. 2000. The *ALS6* and *ALS7* genes of *Candida albicans*. *Yeast* **16:** 847–855.
Hoyer, L.L., Scherer, S., Shatzman, A.R., and Livi, G.P. 1995. *Candida albicans ALS1*: Domains related to a *Saccharomyces cerevisiae* sexual agglutinin separated by a repeating motif. *Mol. Microbiol.* **15:** 39–54.
Hoyer, L.L., Payne, T.L., Bell, M., Myers, A.M., and Scherer, S. 1998a. *Candida albicans ALS3* and insights into the nature of the ALS gene family. *Curr. Genet.* **33:** 451–459.
Hoyer, L.L., Payne, T.L., and Hecht, J.E. 1998b. Identification of *Candida albicans ALS2* and *ALS4* and localization of ALS proteins to the fungal cell surface. *J. Bacteriol.* **180:** 5334–5343.
Hoyer, L.L., Fundyga, R., Hecht, J.E., Kapteyn, J.C., Klis, F.M., and Arnold, J. 2001. Characterization of agglutinin-like sequence genes from non-albicans *Candida* and phylogenetic analysis of the ALS family. *Genetics* **157:** 1555–1567.
Hull, C.M., Raisner, R.M., and Johnson, A.D. 2000. Evidence for mating of the "asexual" yeast *Candida albicans* in a mammalian host. *Science* **289:** 307–310.
Janbon, G., Sherman, F., and Rustchenko, E. 1999. Appearance and properties of L-sorbose-utilizing mutants of *Candida albicans* obtained on a selective plate. *Genetics* **153:** 653–664.
Levinson, G. and Gutman, G.A. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4:** 203–221.
Maxam, A.M. and Gilbert, W. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Meth. Enzymol.* **65:** 499–560.
McNab, R., Holmes, A.R., Clarke, J.M., Tannock, G.W., and Jenkinson, H.F. 1996. Cell surface polypeptide CshA mediates binding of *Streptococcus gordonii* to other oral bacteria and to immobilized fibronectin. *Infect. Immun.* **64:** 4204–4210.
McNab, R., Forbes, H., Handley, P.S., Loach, D.M., Tannock, G.W., and Jenkinson, H.F. 1999. Cell wall-anchored CshA polypeptide (259 kilodaltons) in *Streptococcus gordonii* forms surface fibrils that confer hydrophobic and adhesive properties. *J. Bacteriol.* **181:** 3087–3095.
Moxon, E.R. and Thaler, D.S. 1997. Microbial genetics. The tinkerer's evolving tool-box. *Nature* **387:** 659, 661–652.
Murad, A.M., Leng, P., Straffon, M., Wishart, J., Macaskill, S., MacCallum, D., Schnell, N., Talibi, D., Marechal, D., Tekaia, F., et al. 2001. *NRG1* represses yeast-hypha morphogenesis and hypha-specific gene expression in *Candida albicans*. *EMBO J.* **20:** 4742–4752.
Naglik, J.R., Newport, G., White, T.C., Fernandes-Naglik, L.L., Greenspan, J.S., Greenspan, D., Sweet, S.P., Challacombe, S.J., and Agabian, N. 1999. In vivo analysis of secreted aspartyl proteinase expression in human oral candidiasis. *Infect. Immun.* **67:** 2482–2490.
Nash, T.E. 1997. Antigenic variation in *Giardia lamblia* and the host's immune response. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **352:** 1369–1375.
Odds, F.C. 1988. *Candida and candidosis,* pp. 1–6, 60–230. Bailliere Tindall, London.
Ramasamy, R. 1998. Molecular basis for evasion of host immunity and pathogenesis in malaria. *Biochim. Biophys. Acta* **1406:** 10–27.
Roos, S. and Jonsson, H. 2002. A high-molecular-mass cell-surface protein from *Lactobacillus reuteri* 1063 adheres to mucus components. *Microbiology* **148:** 433–442.
Roy, A., Lu, C.F., Marykwas, D.L., Lipke, P.N., and Kurjan, J. 1991. The *AGA1* product is involved in cell surface attachment of the *Saccharomyces cerevisiae* cell adhesion glycoprotein a-agglutinin. *Mol. Cell. Biol.* **11:** 4196–4206.
Scherer, S. and Stevens, D.A. 1987. Application of DNA typing methods to epidemiology and taxonomy of *Candida* species. *J. Clin. Microbiol.* **25:** 675–679.
Schmid, J., Hunter, P.R., White, G.C., Nand, A.K., and Cannon, R.D. 1995. Physiological traits associated with success of *Candida albicans* strains as commensal colonisers and pathogens. *J. Clin. Microbiol.* **33:** 2920–2926.
Schmid, J., Herd, S., Hunter, P.R., Cannon, R.D., Yasin, M.S.M., Samad, S., Carr, M., Parr, D., McKinney, W., Schousboe, M., et al. 1999. Evidence for a general-purpose genotype in *Candida albicans*, highly prevalent in multiple geographic regions, patient types and types of infection. *Microbiology* **145:** 2405–2414.
Schmitt, M.E., Brown, T.A., and Trumpower, B.L. 1990. A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **18:** 3091–3092.
Schug, M.D., Hutter, C.M., Wetterstrand, K.A., Gaudette, M.S., Mackay, T.F., and Aquadro, C.F. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol. Biol. Evol.* **15:** 1751–1760.
Tibayrenc, M. 1997. Are *Candida albicans* natural populations subdivided? *Trends in Microbiol.* **5:** 253–257.
Vanhee-Brossollet, C. and Vaquero, C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211:** 1–9.
Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M. 1995. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Res.* **23:** 4407–4414.
Welch, D.M. and Meselson, M. 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288:** 1211–1215.
Whelan, W.L. and Soll, D.R. 1982. Mitotic recombination in *Candida albicans*: Recessive lethal alleles linked to a gene required for methionine biosynthesis. *Mol. Gen. Genet.* **187:** 477–485.
Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908–917.
Yang, Z.H. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yang, Z.H. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17:** 32–43.
Zaret, K.S. and Sherman, F. 1982. DNA sequence required for efficient transcription termination in yeast. *Cell* **28:** 563–573.

WEB SITE REFERENCES

http://genome-www4.stanford.edu/cgi-bin/SGD/SAGE/querySAGE; Serial analysis of gene expression data for *S. cerevisiae* at the Stanford Genome Technology Center (SGTC).
http://www-sequence.stanford.edu/group/candida; Sequence data for *C. albicans* strain SC5314 at the SGTC.
http://www-sequence.stanford.edu:8080/bin/blastncontigs6; Sequence data for *C. albicans* strain SC5314 at the SGTC (assembly 6).
http://www-sequence.stanford.edu:8080/contigs/Contig6–2514; Sequence of contig 6–2514 at the SGTC.

Received November 25, 2002; accepted in revised form June 30, 2003.